1 引言

近年来，人们逐渐从信息化时代迈向了数据时代，各种数据爆炸式地增长，数据消费也在日益增多，大量的信息、知识和利润隐藏在这些数据中。如何更有效地利用这些数据，已经成为这个时代下人们共同探索的问题之一。

在这次大作业中，我将对 Adult 数据集进行全面的分析：首先探索数据集中各特征的分布信息；再划分数据集，尝试多种分类模型；最后比较这些模型在 Adult 数据集上的预测结果（分析代码均基于 Python 语言，相关工具和库包可参见附录 A.1）。

2 探索 Adult 数据集

2.1 Adult 数据集的基本信息

Adult 数据集 [1] 也称人口普查收入（Census Income）数据集，来源于美国 1994 年的人口普查数据库，可以作为二分类数据集，用来预测居民年收入是否超过 50K$，其基本信息可参见表 1。

表 1: Adult 数据集的基本信息

属性	值	属性	值
数据集特征	多变量	相关应用	分类
实例数	48842	捐赠日期	1996.5.1
领域	社会	是否有缺失值	有
属性特征	类别型或整数	官网访问次数	1188850
属性数目	14		

Adult 数据集的每个实例包含 14 个属性，其含义、数据类型、取值范围等基本信息见表 2。

表 2: Adult 数据集的基本信息

特征名	含义	数据类型	类别数
age	年龄	整数	-
workclass	工作类型	类别型	8
fnlwgt	序号	整数	-
education	教育程度	类别型	16
education-num	受教育时间	整数	-
marital-status	婚姻状况	类别型	7
occupation	职业	类别型	14
relationship	家庭关系	类别型	6
race	种族	类别型	5
sex	性别	类别型	2
capital-gain	资本收益	整数	-
capital-loss	资本损失	整数	-
hours-per-week	每周工作小时数	整数	-
native-country	原籍	类别型	41

2.2 数据预处理

我首先使用 pandas 库读取 Adult 数据集，将其存储为 pandas 库中的 DataFrame 格式，随机打印出其中几个实例，对该数据集进行初步的观察，结果如下。

```
 1  age  work_class   fnlwgt     education   education_num       marital_status
 2   24     Private   269799    Assoc-voc              11        Never-married
 3   35           ?   169809    Bachelors              13   Married-civ-spouse
 4   51     Private   257126         10th               6   Married-civ-spouse
 5   72     Private   107814      Masters              14        Never-married
 6   33     Private   205950      HS-grad               9        Never-married
 7
 8        occupation     relationship     race     sex   capital_gain
 9   Exec-managerial    Not-in-family    White    Male              0
10                 ?          Husband    White    Male              0
11     Craft-repair           Husband    White    Male              0
12   Prof-specialty    Not-in-family    White    Male           2329
13    Other-service        Own-child    White    Male              0
14
15   capital_loss   hours_per_week   native_country   income
16              0               40    United-States   <=50K.
17              0               20    United-States    <=50K
18              0               40    United-States   <=50K.
19              0               60    United-States    <=50K
20              0               40    United-States    <=50K
```

从以上的初步观察可以得知，Adult 数据集存在数据缺失的情况（如第 3 行和 10 行的“?”），我对整个数据集进行统计后，发现数据集中共有 3620 个实例存在缺失值，而其中 2799 个实例的缺失值多于 1 个（表 3）。同时，我发现分类目标（income）的部分值存在歧义，“<=50K.”与“<=50K”属于同类，却被赋上不同标签，在后续预处理过程中（3.3.3小节）我会进行处理。

表 3: Adult 数据集缺失值分布

	无缺失值	缺失 1 个特征	缺失 2 个特征	缺失 3 个特征
实例数	45222	821	2753	46

考虑到数据集中存在缺失值的实例数较少（仅占总数的 7.41%），且缺失的均为类别型变量，若用一般的方式填补会带来较大的偏差，因此我选择将这些实例直接删除，清理缺失值后的 Adult 数据集包含 45222 个实例，虽然由于删除数据导致 workclass 特征减少了一类（Never-worked），但相应的实例只有 10 个，可以忽略不计。

2.3 Adult 数据集中各特征的分布

对 Adult 数据集进行检查和清理后，我开始探索 Adult 数据集中各特征的分布，我将数值型特征和类别型特征分别处理：对于数值型特征，我主要关注其数字特征（如均值，方差等）及分布密度；对于类别型特征，我主要关注其具体的分布情况。

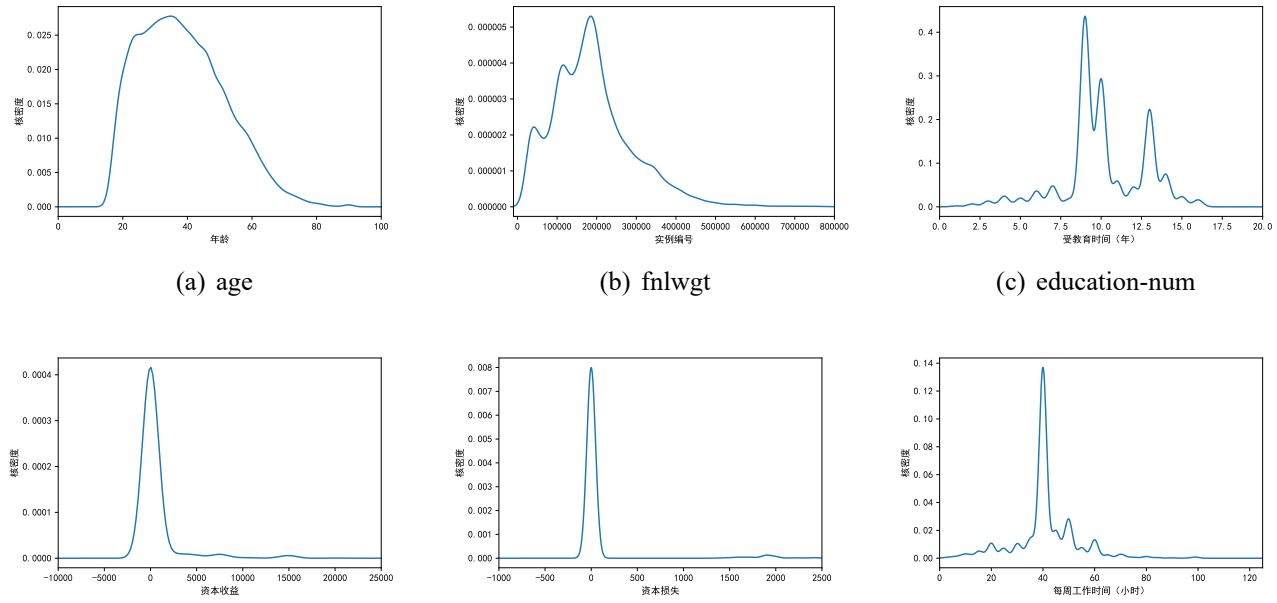
2.3.1 数值型特征的分布

Adult 数据集中的数值型特征为：age, fnlwgt, education-num, capital-gain, capital-loss 以及 hours-per-week。我首先统计其均值、标准差等数字特征，相应结果如表 4所示。

表 4: Adult 数据集数值型特征的数字特征

特征名	均值	标准差	最大值	最小值	上四分位数	下四分位数
age	38.548	13.218	90.000	17.000	47.000	28.000
fnlwgt	18976.470	10563.920	1490400.000	13492.000	237926.000	117388.200
education-num	10.118	2.553	16.000	1.000	13.000	9.000
capital-gain	1101.430	7506.430	99999.000	0.000	0.000	0.000
capital-loss	88.595	404.956	4356.000	0.000	0.000	0.000
hours-per-week	40.938	12.008	99.000	1.000	45.000	40.000

上表的数据大致反映了各特征的分布情况，为更加直观地探索各数值型特征的分布趋势，我作出了相应的概率核密度分布图（高斯核），见图 1。

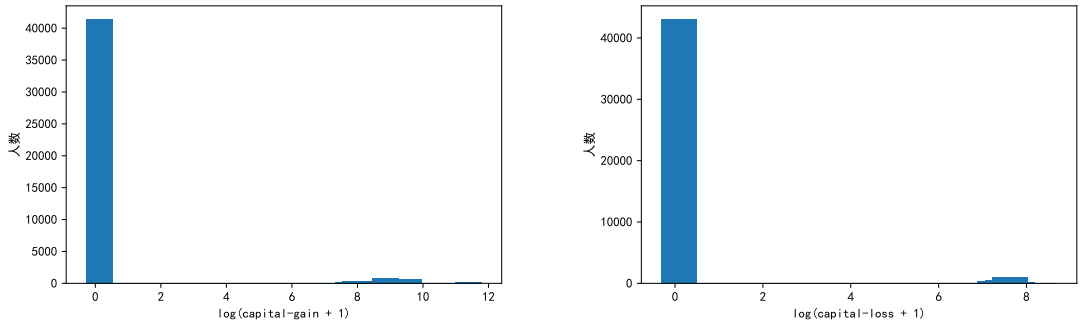


(a) age (b) fnlwgt (c) education-num

(d) capital-gain (e) capital-loss (f) hours-per-week

图 1: Adult 数据集数值型特征概率核密度分布

容易看出，Adult 数据集的 6 个数值型特征接近于正态分布。通过进一步的观察，我发现 capital-gain 和 capital-loss 这两个特征的大部分取值均分布在 0 附近，仅通过概率核密度图无法了解两特征其余取值的分布情况。为更精确、详细地探索其分布，我做出了两特征对数值（$log(x+1)$）下相应的直方图（图 2）。

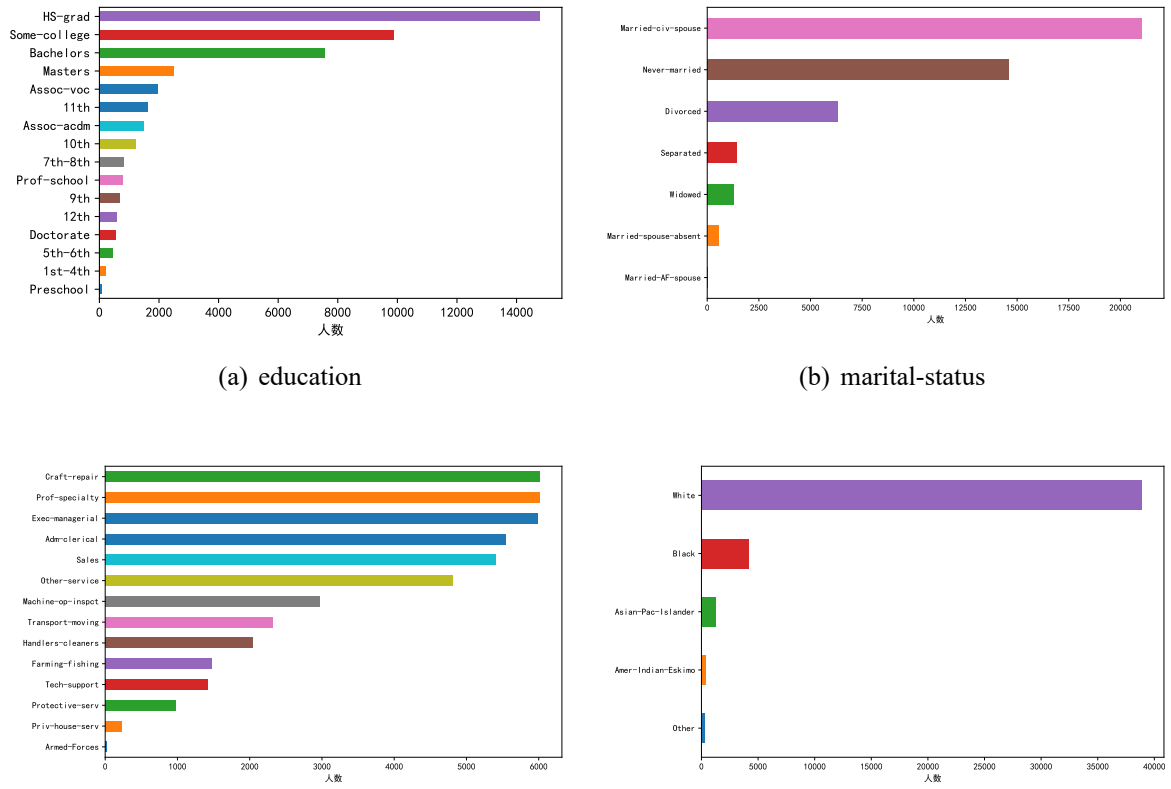


(a) capital-gain (b) capital-loss

图 2: capital-in 和 capital-out 的分布直方图

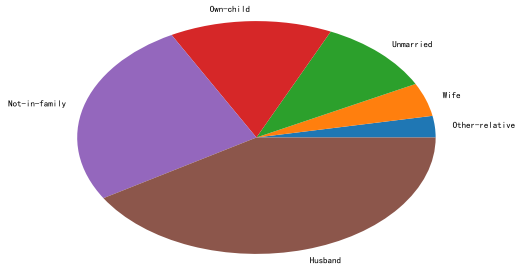
2.3.2 类别型特征的分布

Adult 数据集中的类别型特征包含：workclass, education, marital-status, occupation, relationship, race, sex 以及 native-country。我将其分布表示为条形图或饼图（图 3，4）。各特征中包含的详细类名已记录于附录 A.2.2中。

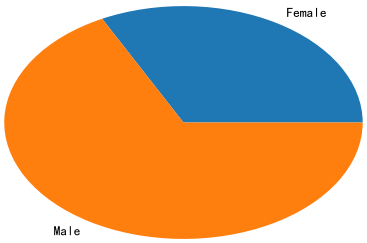


(c) occupation (d) race

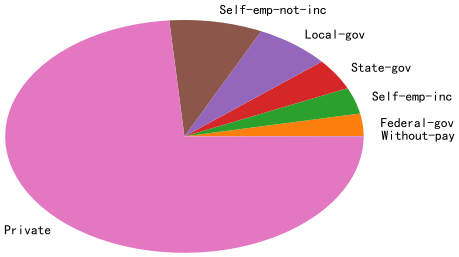
图 3: Adult 数据集类别型特征分布条形图



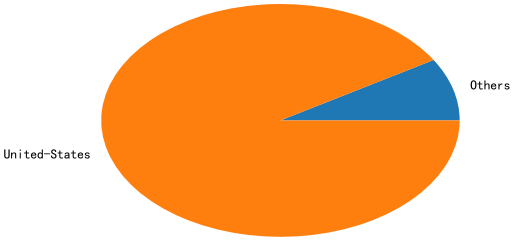
(a) relationship



(b) sex



(c) workclass



(d) workclass

图 4: Adult 数据集类别型特征分布饼图（由于 native-country 包含的类数较多，因此除 United-States 外国籍的分布见附录 A.2.1）

2.4 Adult 数据集中各特征的相关性

探索完 Adult 数据集中各特征的分布后，我开始探索特征之间的相关性，

3 划分数据集并构造分类模型

探索完 Adult 数据集中各特征的分布和相关性后，我开始对其进行训练集和测试集的划分并构造一系列分类模型。

3.1 划分数据集

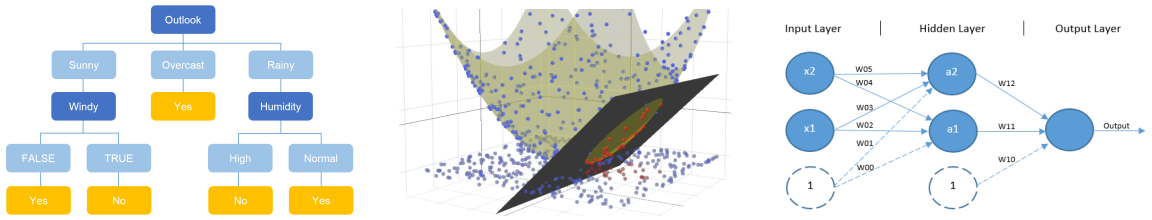
对数据集的划分一般有两种方法：一是直接按照一定的比例将数据划分为训练集和测试集（需保证训练集和测试集中的类分布大致相同）；二是使用分层交叉验证，将数据随机等分为 k 个不相交子集，执行 k 次训练与测试，根据 k 次迭代的平均表现评价模型的性能。

本次作业中，为使评价结果更加精确，我主要使用分层交叉验证方法划分数据集，只在训练基线分类模型（baseline）时使用直接划分训练集和测试集的方法。

3.2 构造分类模型

在机器学习领域，用于分类的算法种类繁多，基本的分类算法包括了逻辑回归（Logistic Regression）、K 近邻（KNN）、决策树、支持向量机（SVM）以及多层感知机（MLP）等。考虑到

Adult 数据集的特征维数并不高，且分类目标简单（二分类），我在本次作业中选择使用决策树、SVM 以及 MLP 三种分类模型（图 5）。除使用单独模型进行分类外，我尝试应用了模型集成的方法以提高相应的分类效果。



(a) 决策树　　(b) SVM　　(c) MLP

图 5: 决策树、SVM 和 MLP 模型的直观表示

3.3 数据预处理

在进行正式的分类之前，我对 Adult 数据集的特征和分类目标进行了一些预处理，以方便分类模型的训练。

3.3.1 Z-score 标准化（规范化）

一般地，Z-score 标准化有如下形式：

$$y = \frac{x - \mu}{\sigma} \tag{1}$$

其中 μ 和 σ 分别代表原数据的均值和标准差。

对于分类问题的数值型特征，经 Z-score 标准化后符合标准正态分布，即 $N(0,1)$，可以有效避免因数值过大导致的模型偏差，并能够加快模型的学习速率，这些优势在 SVM 和 MLP 等模型中表现得更加明显。

除此之外，2.3.1小节的结果表明，Adult 数据集里的数值型特征大多近似服从正态分布，在这个条件下，Z-score 标准化能够取得更良好的效果。若特征的分布与正态分布相差较大，则 Z-score 标准化反而会破坏原数据的分布，造成额外的偏差。

3.3.2 向量化

Adult 数据集中存在 8 个类别型特征，且这些特征的类别之间并无大小关系，如性别的男女之间不存在大小的区别。为方便模型的训练，我将这些特征从字符串转化为 one-hot 向量。经过向量化处理后的数据，每个实例包含 104 维特征。

3.3.3 分类目标修正

Adult 数据集的分类目标为居民收入，分为两类：<=50K$ 以及 >50K$。而数据集中有些实例的分类目标后多了“.”，如变为“<=50K.”，直接使用原数据训练将导致分类目标变为 4 类。因此，我将分类目标转化为-1 和 1，分别表示 <=50K$ 和 >50K$。另外值得注意的一点是 Adult 数据集包含 34014 个负样本（<=50K），而仅有 11208 个正样本。

4 各分类模型的预测结果比较

4.1 基线模型（baseline）

为方便之后的比较，我首先使用 sklearn [8] 中的默认参数，不使用标准化，构造了三个基线模型，以及一个空模型（随机预测），参见表 5和表 6。

表 5: 基线模型在 Adult 数据集上的性能（0.8-0.2 比例的训练集-测试集分割）

	精确度（precision）	召回率（recall）	f1-score	时间（秒）
决策树	0.81	0.81	0.81	0.25
SVM	0.74	0.83	0.76	6.11
MLP	0.78	0.80	0.79	0.59
空模型	0.72	0.52	0.57	0.07

表 6: 基线模型在 Adult 数据集上的性能（5 折分层交叉验证）

	精确度（precision）	召回率（recall）	f1-score	时间（秒）
决策树	0.80	0.80	0.80	4.41
SVM				
MLP	0.81	0.83	0.80	11.82
空模型	0.72	0.50	0.56	0.19

4.2 决策树模型

在本节中，我将使用网格搜索（Grid Search）对决策树模型的分类效果进行评估，搜索的参数及范围参见表 7。

表 7: 对决策树模型网格搜索的参数及范围

参数	含义	类型	范围

最终通过网格搜索筛选出的最佳决策树模型可视化见图 6。

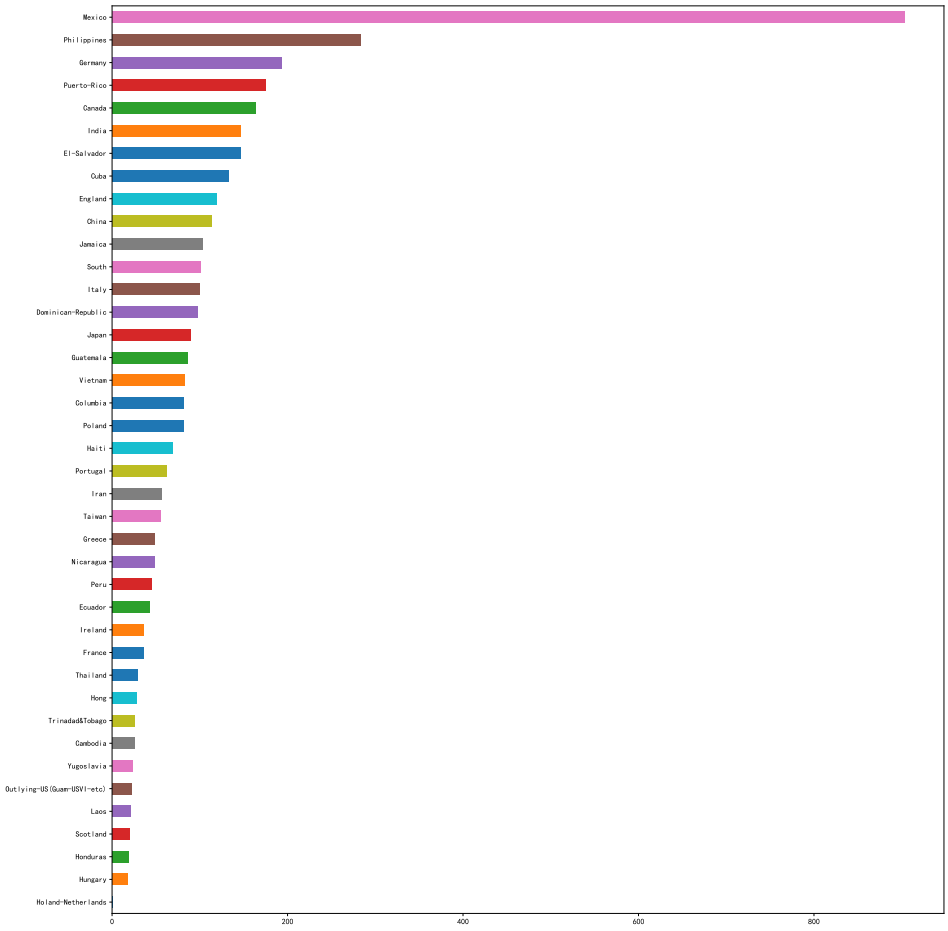


图 6: 最佳决策树模型可视化

4.3 SVM

4.4 MLP

4.5 三种分类模型效果对比

5 对 Adult 数据集的分析结论

Adult 数据集是一个中等规模，存在少部分缺失值的二分类数据集，共 48842 个实例，每个实例包含 14 个特征，其中 8 个为类别型特征，6 个为数值型特征。各类别型特征的分布差异较大，数值型特征的分布近似于正态分布。

A 附录

A.1 作业中使用的工具及库包

本次作业我所使用的编程语言为 Python [2]，编辑环境以 jupyter notebook [3] 为主。作业中我使用的库包见表 8。

表 8: 本作业中使用的库包

库包名	用途
numpy [4]	处理数据，数值计算
pandas [5]	读取数据，绘图，处理数据
matplotlib [6]	绘图
scipy [7]	数值计算
scikit-learn [8]	分类模型的构造和运算

A.2 Adult 数据集特征分布补充资料

A.2.1 native-country 特征的详细分布

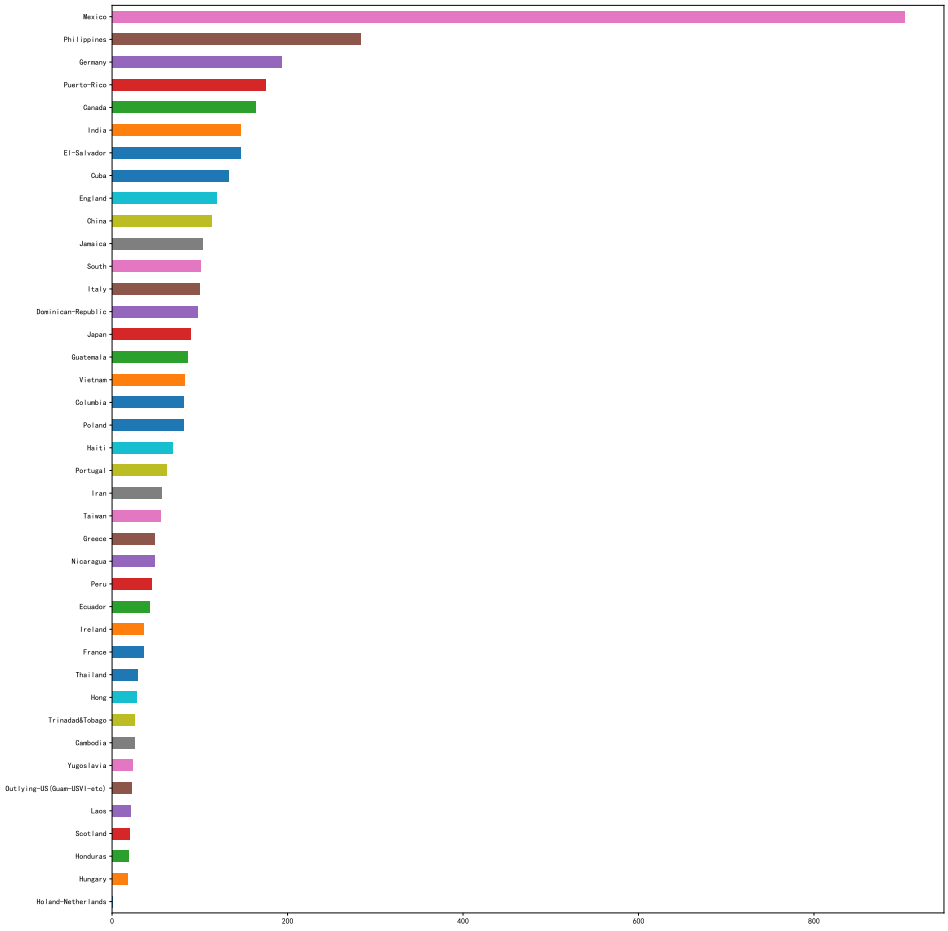


图 7: Adult 数据集 native-country 特征分布（除 United-States）

A.2.2 各类别型特征下的详细类名

1. **workclass** Private, Local-gov, Self-emp-not-inc, Federal-gov, State-gov, Self-emp-inc, Without-pay, Never-worked；
2. **education** 11th, HS-grad, Assoc-acdm, Some-college, 10th, Prof-school, 7th-8th, Bachelors, Masters, Doctorate, 5th-6th, Assoc-voc, 9th, 12th, 1st-4th, Preschool；
3. **marital-status** Never-married, Married-civ-spouse, Widowed, Divorced, Separated, Married-spouse-absent, Married-AF-spouse；
4. **occupation** Machine-op-inspct, Farming-fishing, Protective-serv, Other-service, Prof-specialty, Craft-repair, Adm-clerical, Exec-managerial, Tech-support, Sales, Priv-house-serv, Transport-moving, Handlers-cleaners, Armed-Forces；
5. **relationship** Own-child, Husband, Not-in-family, Unmarried, Wife, Other-relative；
6. **race** Black, White, Asian-Pac-Islander, Other, Amer-Indian-Eskimo；
7. **sex** Male, Female；
8. **native-country** United-States, Cuba, Jamaica, India, Mexico, Puerto-Rico, Honduras, England, Canada, Germany, Iran, Philippines, Poland, Columbia, Cambodia, Thailand, Ecuador, Laos, Taiwan, Haiti, Portugal, Dominican-Republic, El-Salvador, France, Guatemala, Italy, China, South, Japan, Yugoslavia, Peru, Outlying-US(Guam-USVI-etc), Scotland, Trinadad&Tobago, Greece, Nicaragua, Vietnam, Hong, Ireland, Hungary, Holand-Netherlands.

参考文献

[1] D. Dheeru and E. Karra Taniskidou, “UCI machine learning repository,” 2017.

[2] “Python.” https://www.python.org/.

[3] “Jupyter notebook.” http://jupyter.org/.

[4] “Numpy.” http://www.numpy.org/.

[5] “Pandas.” http://pandas.pydata.org/.

[6] “matplotlib.” https://matplotlib.org/.

[7] “scipy.” https://www.scipy.org/.

[8] “scikit-learn.” http://scikit-learn.org/stable/.